Master of Digital Humanities
Introduction to Digital Humanities

Area Maria Guede Ramos - r0915931
areamaria.guederamos@student.kuleuven.be
Jacob Moose - r0871503
jacob.moose@student.kuleveun.be
Annamaria Van Ingelgem - r1008915
annamaria.vaningelgem@student.kuleuven.be

Assignment 1
Short Report
Due Date: 24 November 2023

Table of Contents

Link to our GitHub repository:
https://github.com/MintTeaGreenTea/DH-OpenRefine-Assignment

1. Brief description of the dataset

The dataset contains a series of periodicals published in Belgium between the years 1924 and 2013, although the majority pertain to the early 1980s. The data presented primarily surrounds publication information including the periodicals' city of publication, publication year(s), publication frequency, publisher, and more. Additionally, the dataset provides cataloging numbers and, when available, the physical and digital location of each periodical within KADOC. Most of the periodicals pertain to topics surrounding culture, society, and religion, as expected given KADOC's research focus.

2. Steps taken and choices operated

In order to ensure the safety of our files and to make collaboration possible, we added the files to *GitHub* and created a copy of the dataset. This allowed us to easily refer back to the original data while we cleaned up many of the dataset's columns. In a text document, we documented the steps taken along the way so other team members could easily understand all the changes made. (Key insights from this document are included in section 3 of this report.) Lastly, we included different documents with resources for OpenRefine, GREL, and more, so that these guides would be readily available should we run into any issues while working on the project.

Once the files were secured, we used the MARC 21 Format for Bibliographic Data website to gain a better understanding of what each column and subsection within each column means. This step allowed us to determine which parts of the dataset would prove most useful for the future steps of this project and, more fundamentally, helped us learn as much as possible about the possibilities of OpenRefine.

We decided to focus the majority of our work on cleaning and reconciling columns 245, 260, 310 264, 650, 692, 653, 856, 022, 321, 247, 580, and 775 given the clear publication or bibliographic information they stored. However, before cleaning all these columns, we practiced the cleaning process on other simpler columns that included less or no subsections. Column 310, for example, showcased the publication frequency of each periodical and proved instrumental in familiarizing ourselves with OpenRefine. Following this, we more confidently separated multi-valued cells according to the different occurrences of metadata each cell contained (^$ $ $b, ^$ $ $c, etc.).

For each of the columns we cleaned up, the metadata was removed and relocated to the title itself. Considering the metadata for subsections sequentially followed each of the cells initial string of metadata (i.e., ^$ $ $b following ^$ $ $a ; ^$ $ $c following b, etc.), we did not feel that it was necessary to include all of the metadata in the title, as that would make the final dataset virtually illegible. Irregular metadata that did not follow the standard sequential patterns was either removed and commented on in our "steps taken" document (if suspected data-entry error, for instance) or left untouched.

After our data was fairly clean, we clustered each of the columns in order to get rid of repeated values that slightly differed in spelling, capitalization, or punctuation. For example, "Monthly" was clustered with "monthly" in Column 310 so that the latter would also be capitalized. We repeated this process a few times with the different methods and key functions offered by the software, and we then conducted a manual search using a text facet to find equivalent values that the clustering function missed. For example, the clustering function did not recognize that "10x per jaar" and "Ten times per year" were equivalent, but with the text facet we could easily find them and cluster them. In terms of dates, we left all of them in their original form, though we considered clustering and removing brackets that surrounded many of them.

Clustering the data made us question whether we should normalize and translate the data so it all appeared in English (the dominant language used in the dataset). Considering Anna speaks Dutch, we considered manually replacing each Dutch value using the value.replace() function. Nevertheless, this became rather strenuous and we returned to keeping all values in their original language. In the future, language normalization would play an important role in the data cleaning process.

Lastly, we used the reconcile function in columns 260 and 245 in order to understand how to best make use of it and to link these columns to the already existing Wikidata dataset. This step only proved fruitful in the former column, as it had more general terms ("Leuven" or "Brussels") than the latter, which contained the full names of the periodicals. Additionally, because the WikiData API available within our OpenRefine was only available in English, it proved complicated to reconcile with our information in Dutch, especially since matching would have caused the language of the entry to switch to English. We would have liked to use the Reconcile function for column 650, but given that it was mostly in Dutch, we discarded the idea.

3. Description of the final dataset

Column 260 was split according to the following strings of metadata: "^$ $ $a", "^$ $ $b", "^$ $ $c", "^$ $ $e", and "^$ $ $f". All of the sub-columns in column 260 followed this pattern with the exception of item 739 which included the metadata "$ $ $31981-2015". Based on the column's context, we believe the 3 was an error for "c". Referencing the MARC21 document, we confirmed that the "$ $ $3" value did not make sense in the column and therefore removed it. Column 653 had zero subcolumns, yet still clarified which pieces of data were part of established periodicals. We used value.replace() to remove the "^$ 6a" metadata and made a note in the column's title.

Column 264 was split according to metadata beginning with "^$ 1c" or "^$ 1a". After splitting, we realized that many cells (though not all) included duplicates of dates. We considered removing one set of these dates, as they seemed superfluous, but ended up leaving them both in as not all data had duplicates.

Columns 650, 022, 692, 321, 247, 580 and 775 have been cleaned and renamed. Finally, all columns have been reordered to place the relevant and cleaned columns at the front of the dataset.

In conclusion, we cleaned up the data in order to focus on the most important publication and storage information of the periodicals and to make the dataset fairly legible for any person accessing it.

4. General workflow

To familiarize ourselves with the dataset, Area did the first round of cleaning-up while Jacob and Anna researched the content of some of the columns. After this step, we had a better understanding of the data and what needed to be done. Each team-member was given different columns to clean and tasked with following the procedure described in section 2. We met every week to discuss our work, findings, and issues.

We did several clean-ups independently and experimented with both dividing the columns and simply splitting the multi-valued cells. Through this process, each of us became more familiar with OpenRefine and the many different approaches one can take towards data-cleaning. For instance, at our first meeting we assumed the most productive approach to multi-valued cells would involve splitting subsections into separate columns; however, this action made the dataset harder for us to read when we independently worked on it as more and more columns were created. Returning to the original dataset in our second and third meeting, we focused more on splitting multi-valued cells, moving the metadata to the title, and clustering and reconciling the data.

After creating our final dataset, the short report was worked on and edited by each of the team members. A strong first draft of the methodology was published by Area, and Jacob and Anna followed by adding their individual insights as well as comments on other team members' work. A final revision of both the report and dataset was conducted as a team. A new, cleaner GitHub repository was also created with just the original and final CSV files, so as to avoid the clutter of our original repository.

5. Evaluation of the tool

With reconciling, some matches proved more useful than others. For example, when we began reconciling column 245, the Wikibase instance suggested matching the periodical "Toets" to a wide range of categories, including "keyboard", "pseudoscience", and "test anxiety". None of these classifications made sense with what we knew from the publication, so we did not accept any of these suggestions. Moreover, we determined that the reconcile tool was not useful for the titles section as it simply got rid of the original title and substituted it for the entry of the related entry found on Wikidata; for example, for the publication "Onze lieve vrouw van Kortenbos", the Reconcile function simply changed this title for "Basiliek van Onze-Lieve-Vrouw van Kortenbos", to which it's clearly related, but the two are not equivalent. Had it kept the original title while providing the reference link to the

church, the reconciliation would have proved useful, but in its substitution it caused us to lose specificity and important data.

However, in other instances - such as the names of cities - the tool worked successfully (though it took a fairly long time to operate). Because only specific types of information could be reconciled and we chose to split columns into multi-valued cells, we ran into quite a few difficulties with the reconciliation. For example in column 260 we could reconcile the place of publication and sometimes the organization from which the publication originated, but the years of publication were never reconcilable. Because we split this column into multi-valued cells and didn't export the place of publication to another column, it was very cumbersome to reconcile the entire column, as most information was not reconcilable. This issue made it so difficult and time-consuming to go through all the possible matches, we decided not to continue reconciling this column. Nevertheless, this attempt helped us in better understanding the scope and limitations of the Reconcile tool.

OpenRefine generally proved incredibly useful in the clean up of the data, as it allowed us to easily replace metadata for a large number of cells. Moreover, the different approaches available - splitting data by columns, rows, cells themselves, etc. - provided us with many choices and opportunities to clean our database. Having an undo button that allowed us to go back and revise mistakes was also invaluable to our project. Nevertheless, we wish there was an easier way the data could be accessed by multiple users concurrently. This would have made it much easier for our group to work together at the same time.

6. Examples of GREL used and of their utility

The value.replace('value1', 'value2') function was useful for removing metadata, especially the initial strings of metadata that cells often started with (i.e., "$ $ $a"). Other expressions that we experimented with included replacing only the first "^" with a "" and separating each cell based on proceeding "^" values. Additionally, we noticed that the "^" value or "$" symbol were key indicators of where we could use the built-in split function.

We experimented with making use of the "startsWith()" and "endsWith()" functions to find particular items that started or ended with certain characters within our dataset, as well as with "toNumbers()" to change strings into numbers, but we did so in order to further familiarize ourselves with the programs and the functions, and we ended up discarding the changes made with these functions. Ultimately, we performed the clean-up by using "replace()"—as it quickly became clear that it was the most efficient function for our task—, as well as the built-in functions in OpenRefine, such as "Split multi-valued cells", "Transform", and "Cluster and edit", and the multiple facets that allowed us to find outliers.